

Galaxy for Data Provenance

Tom Doak

Le-Shin Wu

Carrie Ganote

National Center for Genome Analysis Support

July 16, 2014



INDIANA UNIVERSITY



INDIANA UNIVERSITY

Galaxy for Data Provenance

So...

What does data provenance mean?

The ability to reproduce and validate the scientific methods and outcomes given sufficient documentation of the data



INDIANA UNIVERSITY

Why is this important?



WOW, this unicorn mantis has never been seen by mankind!

I better get my camera...



INDIANA UNIVERSITY

Why is this important?

Without sufficient evidence to be able to reproduce your research, the conclusions are less compelling



NOOOOOOOOOOOOOOOOOOOOO
My dissertation!!!!



What needs to be considered?

- Raw data – how was it generated
- What software was used at each step (version, environment, operating system)
- What parameters were used
- How were intermediate files handled (copied, renamed, deleted)
- Backtracking, param sweeps, etc



How is provenance handled?

Common practices for analysis tracking involve:

- Lab Notebooks – paper and pen
- Naming files and directories descriptively
- Readme type files with each step documented
- Wiki pages for tracking steps
- Workflow-specific software
- Complicated custom solutions involving hashing everything




Pitfalls of these approaches

- Loss and damage potential
- They depend on the detail of the researcher
- Can involve more work than the actual work
- They may fail to make sense when revisited a year+ later



Does this look familiar?

Name ▼	Date Modified	Date Created
▶ visuals	3/12/13, 2:43 PM	Feb 11, 2013
▶ TripleChecking	1/30/13, 3:30 PM	Jan 30, 2013
▶ TrackedOnline	1/25/13, 5:37 PM	Oct 26, 2012
▶ Temperature	4/30/13, 3:48 PM	Nov 19, 2012
▶ Raw_Inputs	6/22/14, 7:44 PM	Apr 30, 2013
 R-code-fo...p-covars.r	2/14/13, 5:12 PM	Feb 14, 2013
▶ partek	2/21/13, 5:34 PM	Feb 13, 2013
▶ Papers	7/8/14, 11:49 AM	Apr 30, 2013
▶ Oct2013	11/12/13, 11:21 PM	Oct 31, 2013
▶ Learnings	4/9/13, 3:02 PM	Nov 16, 2012
▶ Deadends	5/1/13, 4:59 PM	Jan 25, 2013
▶ August2013	8/21/13, 3:03 PM	Aug 21, 2013
▶ Analysis	6/23/14, 11:02 AM	Dec 4, 2012,



Does this look familiar?

Adequate documentation
practices are difficult to teach
and exercise

```
addTagToFastqHeaders.pl_042414.11:33:28.temp  
addTagToFastqHeaders.pl_042514.16:02:37.temp  
addTagToFastqHeaders.pl_042514.16:24:18.temp  
addTagToFastqHeaders.pl042814_042814.16:39:27.temp  
addTagToFastqHeadersSkipmismatches.pl_050714.17:02:27.temp  
addTagToFastqHeadersSkipmismatches.pl_050714.17:07:05.temp  
addTagToFastqHeadersSkipmismatches.pl_050714.17:08:27.temp  
addTagToFastqHeadersSkipmismatches.pl_050714.17:09:58.temp  
addTagToFastqHeadersSkipmismatches.pl_050714.17:13:38.temp  
addTagToFastqHeadersSkipmismatches.pl_050714.17:17:25.temp
```

```
addTagToFastqHeaders.pl  
alterFastq.pl~  
checkFastqIntegrity_calculateAfter.pl  
checkFastqIntegrity_calculateAfter.pl~  
checkFastQIntegrity.pl  
checkFastQIntegrity.pl~  
checkFastQIntegrity010214_recovered.pl  
checkFastQIntegrity010214.pl  
checkFastQIntegrity010214.pl~  
checkFastQIntegrity042214.pl  
checkFastQIntegrity042214.pl~  
fakeread_left.fastq
```



INDIANA UNIVERSITY

Version Control?

We can't just
use Github

Great for tracking changes
in a code base, not great
for tracking a genomics
project and parameter
sweeping





How does Galaxy help?

- History tracks every change made to a file
- When copying files to a new history/ sharing files with other users, hints about their origin are generated
- Galaxy uses version control to keep track of the exact tool wrapper used
- New features in Galaxy allow users to specify which version of the software the tool uses

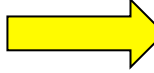
The screenshot shows the 'History' panel in the Galaxy web interface. At the top, it says 'Tuxedo Suite - nematode' with a size of '8.1 GB'. Below this is a list of workflow steps, each with a description and icons for viewing, editing, and deleting. The steps are numbered 151 through 159. Steps 156, 157, 158, and 159 are highlighted in green, indicating they are the most recent or active steps. The descriptions for these steps are: 'cummeRbund on data 149, data 148, and others (HTML)', 'cummeRbund on data 149, data 148, and others: Database File (sqlite)', 'cummeRbund on data 149, data 148, and others (HTML)', and 'cummeRbund on data 149, data 148, and others: Database File (sqlite)'. Steps 151, 152, and 153 are highlighted in light green and describe 'Cuffdiff for cummeRbund on data 54, data 44, and others: transcript FPKM tracking', 'Cuffdiff for cummeRbund on data 54, data 44, and others: transcript differential expression testing', and 'Cuffdiff for cummeRbund on data 54, data 44, and others: gene'.

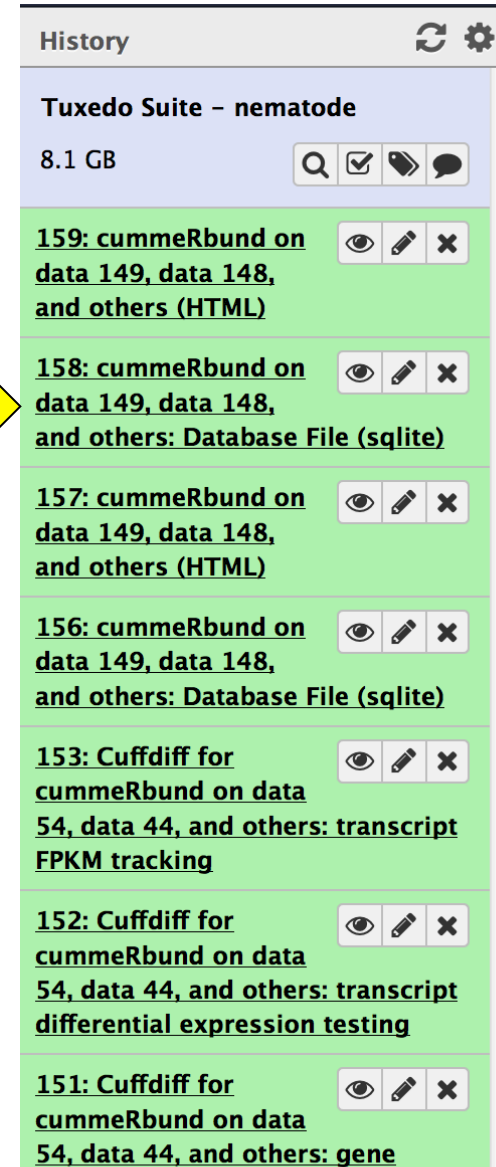
Step Number	Description	View	Edit	Delete
159	cummeRbund on data 149, data 148, and others (HTML)			
158	cummeRbund on data 149, data 148, and others: Database File (sqlite)			
157	cummeRbund on data 149, data 148, and others (HTML)			
156	cummeRbund on data 149, data 148, and others: Database File (sqlite)			
153	Cuffdiff for cummeRbund on data 54, data 44, and others: transcript FPKM tracking			
152	Cuffdiff for cummeRbund on data 54, data 44, and others: transcript differential expression testing			
151	Cuffdiff for cummeRbund on data 54, data 44, and others: gene			









Shortcomings of Galaxy




- Generated history names are hard to parse
- Under normal settings, user may not know for sure what version of (software, database, etc) they have used
- Does not prevent human error – renaming wrong file, admin error, etc.




What? 









History  




Tuxedo Suite – nematode
8.1 GB    




159: cummeRbund on data 149, data 148, and others (HTML)   




158: cummeRbund on data 149, data 148, and others: Database File (sqlite)   

157: cummeRbund on data 149, data 148, and others (HTML)   

156: cummeRbund on data 149, data 148, and others: Database File (sqlite)   

153: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript FPKM tracking   

152: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript differential expression testing   

151: Cuffdiff for cummeRbund on data 54, data 44, and others: gene   



Adding more information

Renaming the history items makes a history much cleaner to follow – tracking back the datasets by number is not ideal

Galaxy allows annotation and note-taking per dataset in the history

The screenshot displays the Galaxy web interface. At the top is a navigation bar with links: Analyze Data, Workflow, Shared Data, Visualization, Admin, Help, and User. Below this is a sub-navigation bar with tabs: Attributes, Convert Format, Datatype, and Permissions. The 'Attributes' tab is active, showing the 'Edit Attributes' dialog. This dialog has three sections: 'Name:' with a text input field containing 'cummeRbund on merged salt treated'; 'Info:' with a text input field containing 'CuffSet instance with: 2 samples 14758 genes'; and 'Annotation / Notes:' with a large text area containing a detailed description of the experiment. To the right of the dialog is the 'History' panel, which lists several history items. Each item has a title, a size indicator, and icons for viewing, editing, and deleting. The items are: 159: cummeRbund on data 149, data 148, and others (HTML); 158: cummeRbund on data 149, data 148, and others: Database File (sqlite); 157: cummeRbund on data 149, data 148, and others (HTML); 156: cummeRbund on data 149, data 148, and others: Database File (sqlite); 153: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript FPKM tracking; 152: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript differential expression testing; and 151: Cuffdiff for cummeRbund on data 54, data 44, and others: gene.

Attributes Convert Format Datatype Permissions

Edit Attributes

Name:
cummeRbund on merged salt treated

Info:
CuffSet instance with:
2 samples
14758 genes

Annotation / Notes:
The samples from the salt treated set were incubated in a .15% saline solution for 3 days at a temperature of 30 degrees. They were then cooked lightly in ketchup before undergoing library prep using the TruSeq blahblahblah kit. They were sequenced at Tufts at an approximate depth of 50x with read lengths of 100bp, paired end. They were run on 2 lanes of an Illumina HiSeq blahblah machine. They then underwent basic adapter cutting and quality control from the sequencing center.

Add an annotation or notes to a dataset; annotations are available when a history is viewed.

History

Tuxedo Suite – nematode
8.1 GB

159: cummeRbund on data 149, data 148, and others (HTML) Edit attributes

158: cummeRbund on data 149, data 148, and others: Database File (sqlite)

157: cummeRbund on data 149, data 148, and others (HTML)

156: cummeRbund on data 149, data 148, and others: Database File (sqlite)

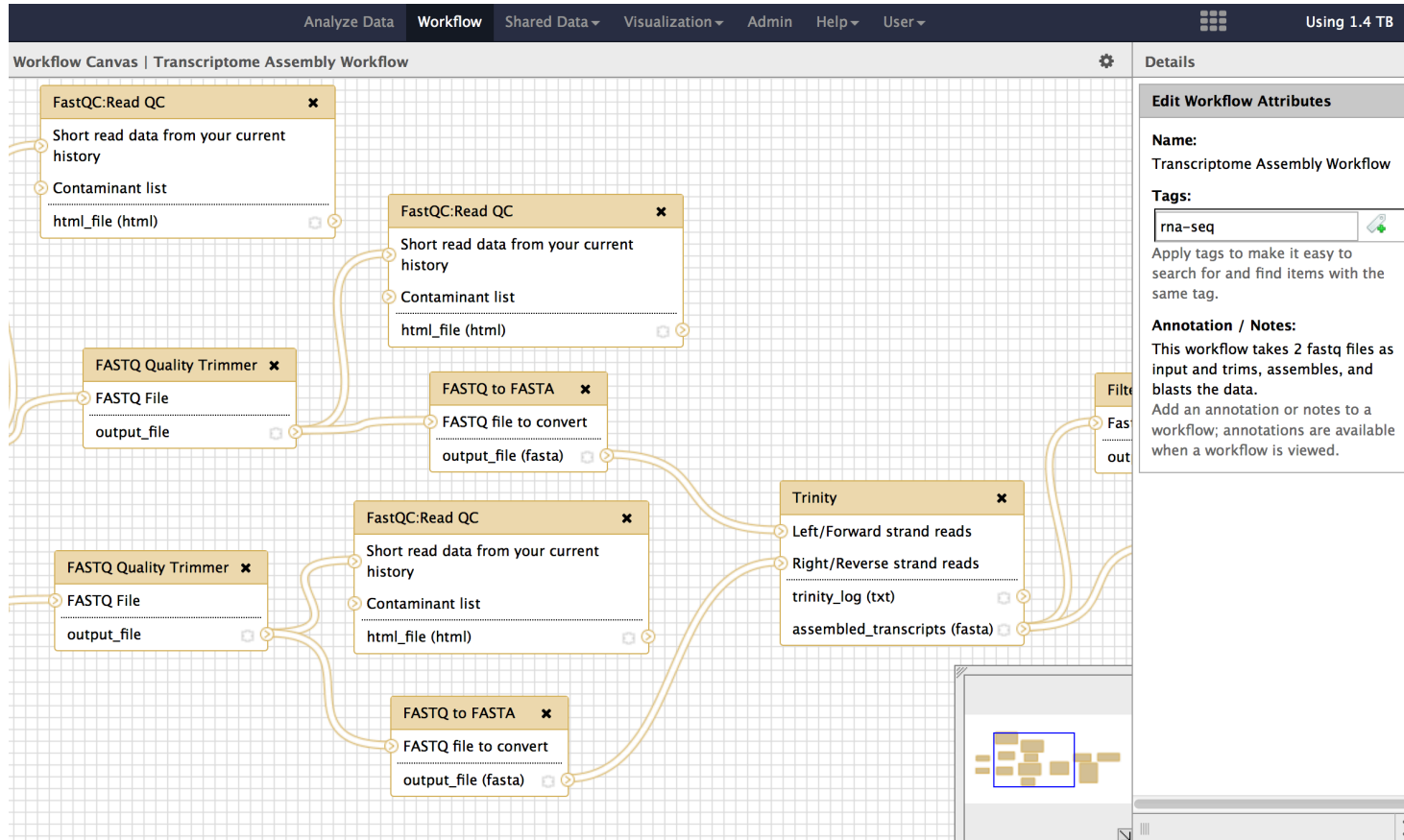
153: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript FPKM tracking

152: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript differential expression testing

151: Cuffdiff for cummeRbund on data 54, data 44, and others: gene



Adding more information



Notes
can be
added to
workflows
as well as
histories



Adding more information

Individual steps can be renamed at runtime, as well.

The screenshot displays a workflow editor interface. On the left, a grid contains several workflow steps. One step, labeled 'FASTQ to FASTA', is highlighted with a blue box. This step has two inputs: 'FASTQ file to convert' and 'output_file (fasta)'. To the right of the grid is a 'Details' panel for the selected step. This panel includes the tool name 'FASTQ to FASTA', version '1.0.0', and a description of the step's function. Below this, there are sections for 'Edit Step Actions' (containing a 'Rename Dataset' dropdown and a 'Create' button) and 'Edit Step Attributes' (containing an 'Annotation / Notes' text area). At the bottom of the details panel is a 'Citation' section. The top of the interface shows a user profile and a storage indicator 'Using 1.4 TB'.

ser ▾ Using 1.4 TB

Details

Tool: FASTQ to FASTA

Version: 1.0.0

FASTQ file to convert
Data input 'input_file' (fastq)

Edit Step Actions

Rename Dataset ▾

output_file ▾ Create

Add actions to this step; actions are applied when this workflow step completes.

Edit Step Attributes

Annotation / Notes:

Add an annotation or notes to this step; annotations are available when a workflow is viewed.

What it does

This tool converts FASTQ sequencing reads to FASTA sequences.

Citation



Backtracking

Items hidden/
deleted from the
history are not
gone forever –
encouraging
pruning of
histories to
retain the
current
workflow

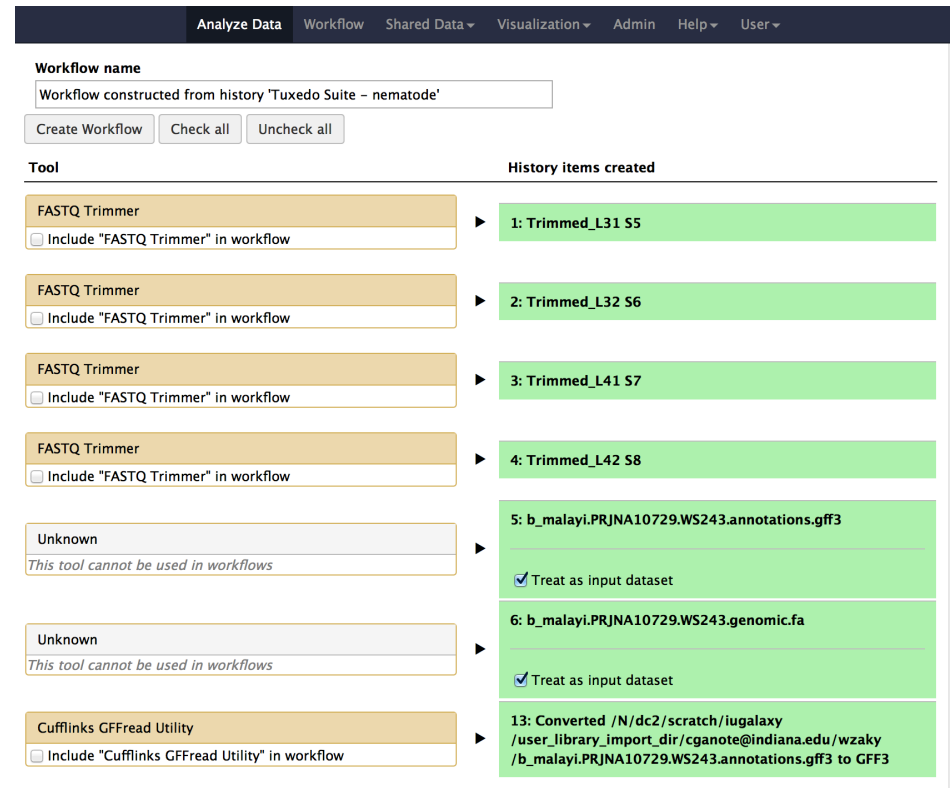
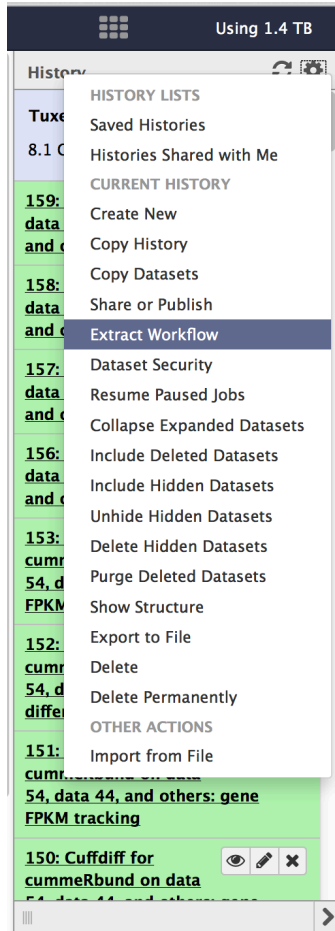
The screenshot shows the 'History' tab in the Indiana University Genomics Data Commons interface. The top navigation bar includes 'Workflow', 'Shared Data', 'Visualization', 'Admin', 'Help', and 'User'. The right side of the bar shows 'Using 1.4 TB'. The main content area displays a list of history items, each with a status icon (yellow triangle for deleted, red X for hidden) and a description. A context menu is open on the right side of the list, showing options for managing the history. The menu includes 'HISTORY LISTS', 'Saved Histories', 'Histories Shared with Me', 'CURRENT HISTORY', 'Create New', 'Copy History', 'Copy Datasets', 'Share or Publish', 'Extract Workflow', 'Dataset Security', 'Resume Paused Jobs', 'Collapse Expanded Datasets', 'Include Deleted Datasets' (which is highlighted), 'Include Hidden Datasets', 'Unhide Hidden Datasets', 'Delete Hidden Datasets', 'Purge Deleted Datasets', 'Show Structure', 'Export to File', 'Delete', and 'Delete Permanently'.

History Item	Status
156: cummeRbund on data 149, data 148, and others: Database File (sqlite)	Deleted
155: cummeRbund on data 149, data 148, and others (HTML)	Hidden
154: cummeRbund on data 149, data 148, and others: Database File (sqlite)	Deleted
153: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript FPKM tracking	Visible
152: Cuffdiff for cummeRbund on data 54, data 44, and others: transcript differential expression	Visible
151: Cuffdiff for cummeRbund on data 54, data 44, and others: gene FPKM tracking	Visible
150: Cuffdiff for cummeRbund on data 54, data 44, and others: gene differential expression	Visible
149: Cuffdiff for cummeRbund on data 54, data 44, and others: TSS groups FPKM tracking	Visible
148: Cuffdiff for cummeRbund on data 54, data 44, and others: TSS groups differential expression	Visible
147: Cuffdiff for cummeRbund on data 54, data 44, and others: CDS FPKM tracking	Visible
146: Cuffdiff for cummeRbund on data 54, data 44, and others: CDS FPKM differential expression	Visible



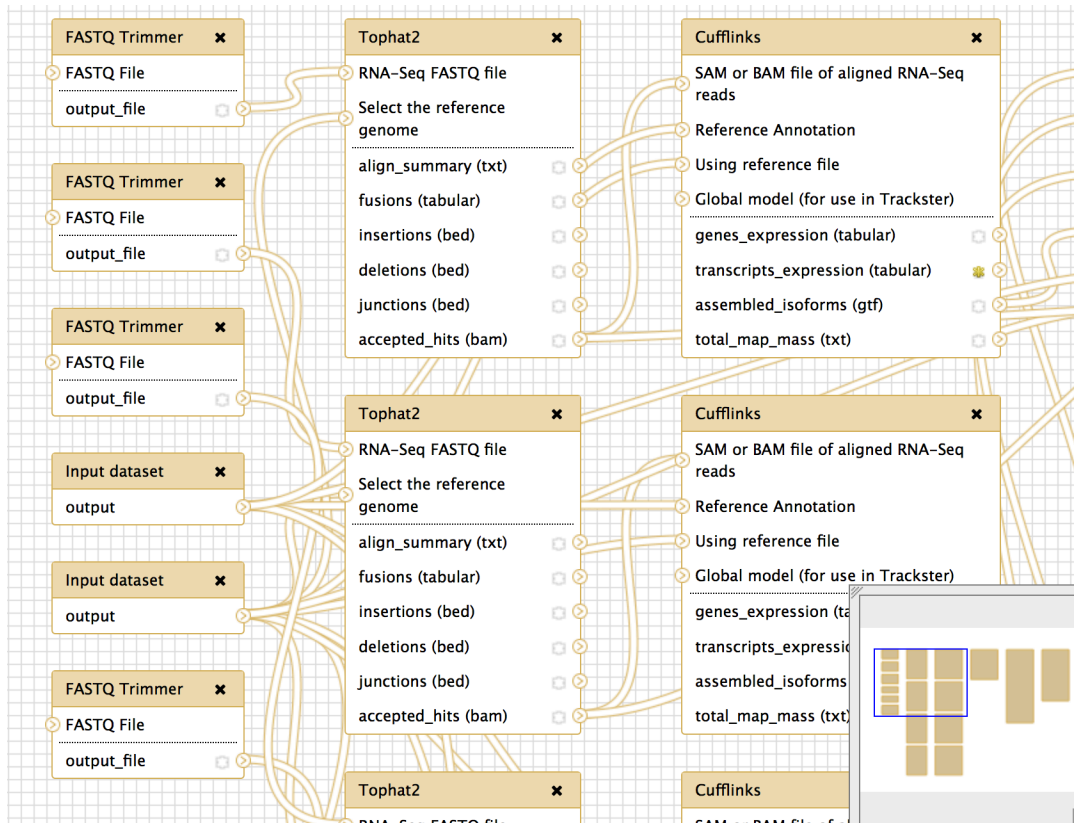
Visualizing Histories

Creating a workflow from a history is one way to visualize the pipeline, and also to ensure that the pipeline works as intended





Visualizing Histories



Pipelines can be quite complex – but if you are good at reading spaghetti, then this is the method for you



Provenance – how much is too much?

It is good to know and record, for example, which release of NR you ran your blast job on

Keep a copy of every release for sake of reproducing blast results?

We keep 121GB of Blast database available on DC2 and update monthly

What about obsolete software?

Different results may be derived from different machines – decommissioned machines are long gone

Planetary alignments?



End Goals

Galaxy workflows and raw inputs are enough to publish a perfectly reproducible experiment

The same results 1, 5 and 10 years from now

We are not quite there yet!



INDIANA UNIVERSITY

Fin

Thanks for watching!
Questions and comments:
Email help@ncgas.org